APPLICATION OF LOGISTIC REGRESSION MODEL AND ITS VALIDATION FOR LANDSLIDE SUSCEPTIBILITY MAPPING USING GIS AND REMOTE SENSING DATA AT PENANG, MALAYSIA

SARO. LEE
Geoscience Information Center, Korea Institute of Geology & Mineral Resources (KIGAM)
30, Gajung-dong, Yusung-gu, Daejeon, 305-350, Korea
leesaro@kigam.re.kr

Abstract: The aim of this study is to evaluate the hazard of landslides at Penang, Malaysia, using a Geographic Information System (GIS) and remote sensing. Landslide locations were identified in the study area from interpretation of aerial photographs and from field surveys. Topographical and geological data and satellite images were collected, processed, and constructed into a spatial database using GIS and image processing. The factors chosen that influence landslide occurrence were: topographic slope, topographic aspect, topographic curvature and distance from drainage, all from the topographic database; lithology and distance from lineament, taken from the geologic database; land use from TM satellite images; and the vegetation index value from SPOT satellite images. Landslide hazardous area were analysed and mapped using the landslide-occurrence factors by logistic regression model. The results of the analysis were verified using the landslide location data and compared with probabilistic model. The validation results showed that the logistic regression model is better prediction accuracy than probabilistic model.

Keywords: Landslide; susceptibility; logistic regression; GIS; Remote Sensing; Penang

1. INTRODUCTION

Recently there has been an increasing occurrence of landslides in Malaysia. Most of these landslides occurred on cut slopes or on embankments alongside roads and highways in mountainous areas. Some of these landslides occurred near high-rise apartments and in residential areas, causing great anxiety in many people. A few major and catastrophic landslides have also occurred within the last 10 years. These landslides have resulted in significant damage to both people and property. In the area chosen in this study, Penang in Malaysia, much damage was caused on each of these occasions. The trigger for the landslides was a period of heavy rainfall, and, as there was little effort to assess or predict the event, damage was extensive. Through scientific analysis of landslides, we can assess and predict landslide-susceptible areas, and thus decrease landslide damage through proper preparation. To achieve this aim, landslide susceptibility analysis techniques have been applied, and verified in the study area. In addition, landslide-related factors were also assessed.

The Penang area has suffered much landslide damage following heavy rains, and was selected as a suitable candidate to evaluate the frequency and distribution of landslides. Penang is one of the 13 states of the Federation of Malaysia. The Penang area is located on the northwest coast of the Malaysian peninsular. It is bounded to the north and east by the state of Kedah, to the south by the state of Perak, and to the west by the Straits of Malacca and by Sumatra (Indonesia). Penang consists of the island of Penang, and a coastal strip on the mainland, known as Province Wellesley. The island covers an area of 285 km^2, and is separated from the mainland by a channel. The rainfall is quite evenly distributed throughout the year, with more rain occurring from September to November. Penang has a population of approximately one million people. The bedrock geology of the study area consists mainly of granite.

There have been many studies carried out on landslide hazard evaluation using GIS; for example, Guzzetti *et al.*. (1999) summarized many landslide hazard evaluation studies. Recently, there have been studies on landslide hazard evaluation using GIS, and many of these studies have applied probabilistic methods. One of the statistical methods available, the logistic regression method, has also been applied to landslide hazard mapping. As a new approach to landslide hazard evaluation using GIS, data mining using fuzzy logic, and artificial neural network methods have been applied.

A key assumption using this approach is that the potential (occurrence possibility) of landslides will be comparable to the actual frequency of landslides. Landslide occurrence areas were detected in the Penang area, Malaysia by interpretation of aerial photographs and field surveys. A map of landslides was developed from aerial photographs, in combination with the GIS, and this were used to evaluate the frequency and distribution of shallow landslides in the area. Topography and lithology databases were constructed and lineament, land use and vegetation index value extracted from Landsat TM and SPOT XS satellite image for the analysis. Then, the calculated and extracted factors were converted to a 10m × 10m grid (ARC/INFO GRID type). Then, using logistic regression, one of the statistical models, the landslide occurrence possibility was derived as formula. The formula was used for calculating the landslide susceptibility index, and the index was mapped to represent landslide susceptibility. Finally, the susceptibility map was verified using known landslide locations and success rates were calculated for quantitative validation. To compare the validation result, probabilistic model, frequency ratio was applied using the same database. Using the frequency ratio models, the spatial relationships between the landslide location and each landslide related factor

was extracted using the relationships. Then, the relationship was used as each factor's rating in the overlay analysis In the study, Geographic Information System (GIS) software, ArcView 3.2, and ARC/INFO 8.1 NT version software packages were used as the basic analysis tools for spatial management and data manipulation.

2. DATA USING GIS AND REMOTE SENSING

Accurate detection of the location of landslides is very important for probabilistic landslide susceptibility analysis. The application of remote sensing methods, such as aerial photographs and satellite images, are used to obtain significant and cost-effective information on landslides. In this study, 1:10,000–1:50,000-scale aerial photographs were used to detect the landslide locations. These photographs were taken during the period 1981–2000, and the landslide locations were detected by photo interpretation and the locations verified by fieldwork. Recent landslides were observed in aerial photographs from breaks in the forest canopy, bare soil, or other geomorphic characteristics typical of landslide scars, for example, head and side scarps, flow tracks, and soil and debris deposits below a scar. To assemble a database to assess the surface area and number of landslides in each of three study areas, a total of 541 landslides were mapped in a mapped area of 293 km^2.

Identification and mapping of a suitable set of instability factors having a relationship with the slope failures requires an *a priori* knowledge of the main causes of landslides (Guzzetti *et al.* 1999). These instability factors include surface and bedrock lithology and structure, bedding altitude, seismicity, slope steepness and morphology, stream evolution, groundwater conditions, climate, vegetation cover, land use, and human activity. The availability of thematic data varies widely, depending on the type, scale, and method of data acquisition. To apply the probabilistic method, a spatial database that considers landslide-related factors was designed and constructed. These data are available in Malaysia either as paper or as digital maps.

There were eight factors considered in calculating the probability, and the factors were extracted from the constructed spatial database. The factors were transformed into a vector-type spatial database using the GIS, and landslide-related factors were extracted using our database. Using the topographic database, a digital elevation model (DEM) was created first. Contour and survey base points that had elevation values read from the 1:50,000-scale topographic maps were extracted, and a DEM was constructed with a resolution of 10 m. Using this DEM, the slope angle, slope aspect, and slope curvature were calculated. In the case of the curvature negative curvatures represent concave, zero curvature represent flat and positive curvatures represents convex. In addition, the distance from drainage was calculated using the topographic database. The drainage buffer was calculated in 100 m intervals. Using the geology database, the lithology was extracted, and the distance from lineament calculated. The lithology map obtained from a 1:50,000-scale geological map. The lineament buffer was calculated in 100 m intervals. Land use data was classified using a LANDSAT TM image employing an unsupervised classification method and field survey. The 11 classes identified, such as urban, water, forest, agricultural area, and barren area were extracted for land use mapping. Finally, the Normalized Difference Vegetation Index (NDVI) was obtained from SPOT satellite. The NDVI value was calculated using the formula NDVI = (IR – R)/(IR + R), where IR value is the infrared portion of the electromagnetic spectrum, and R-value is the red portion of the electromagnetic spectrum. The NDVI value denotes areas of vegetation in an image.

The factors were converted to a raster grid with 10 m × 10 m cells for application of the logistic regression and frequency ratio model. The area grid was 2,490 rows by 1,884 columns (i.e., total number is 4,691,160) and 541 cells had landslide occurrence.

3. LOGISTIC REGRESSION MODEL AND ITS APPLICATION

Logistic regression allows one to form a multivariate regression relation between a dependent variable and several independent variables. Logistic regression, which is one of the multivariate analysis models, is useful for predicting the presence or absence of a characteristic or outcome based on values of a set of predictor variables. The advantage of logistic regression is that, through the addition of an appropriate link function to the usual linear regression model, the variables may be either continuous or discrete, or any combination of both types and they do not necessarily have normal distributions. In the case of multi-regression analysis, the factors must be numerical, and in the case of a similar statistical model, determinant analysis, the variables must have a normal distribution. In the present situation, the dependent variable is a binary variable representing presence or absence of landslide. Where the dependent variable is binary, the logistic link function is applicable (Atkinson and Massari, 1998). For this study, the dependent variable must be input as either 0 or 1, so the model applies well to landslide possibility analysis. Logistic regression coefficients can be used to estimate ratios for each of the independent variables in the model.

Quantitatively, the relationship between the occurrence and its dependency on several variables can be expressed as:

$$p = 1 / (1 + e^{-z}) \quad (1)$$

where p is the probability of an event occurring. In the present situation, the value p is the estimated probability of landslide occurrence. The probability varies from 0 to 1 on an S-shaped curve and z is the linear combination. It follows that logistic regression involves fitting an equation of the following form to the data:

$$z = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n \quad (2)$$

where b_0 is the intercept of the model, the b_i (i = 0, 1, 2, ..., n) are the slope coefficients of the logistic regression model, and the x_i (i = 0, 1, 2, ..., n) are the independent variables. The linear model formed is then a logistic regression of presence or absence of landslides (present conditions) on the independent variables (pre-failure conditions).

Using the logistic regression model, the spatial relationship between landslide-occurrence and factors influencing landslides was assessed. The spatial databases of each factor were converted to ASCII format files for use in the statistical package, and the correlations between landslide and each factor were calculated. There are two cases. A first case is the only one factor was used. In this case, logistic regression formulae were created as shown in equations from (3) to (9) for each case. Finally, the probability that predicts the possibility of landslide-occurrence was calculated using the spatial database, data and equations (1) and from (3) to (9). A second case is the all factor were used. In this case, logistic regression formulae were created as shown in equations (1) and (10) for each case.

$$z_1 = (0.0399 \times \text{SLOPE}) - 9.2011 \quad (3)$$

$$z_2 = (9.12\text{E-}05 \times \text{CURVATURE}) - 8.5962 \quad (4)$$

$$z_3 = \text{ASPECT}_c - 8.6546 \quad (5)$$

$$z_4 = (-.0011 \times \text{DRAINAGE}) - 8.4254 \quad (6)$$

$$z_5 = \text{LITHOLOGY}_c - 8.3756 \quad (7)$$

$$z_6 = (-.0009 \times \text{FAULT}) - 8.0873 \quad (8)$$

$$z_7 = \text{LANDUSE}_c - 12.2029 \quad (9)$$

$$z_8 = (0086 \times \text{NDVI}) - 8.8239 \quad (10)$$

$$z_9 = (0.0308 \times \text{SLOPE}) + \text{ASPECT} + (-0.0000512 \times \text{CURVATRURE}) + (0.0006 \times \text{DRAINAGE}) + \text{LITHOLOGY}_b + (-0.0012 \times \text{FAULT}) + (-0.0053 \times \text{NDVI})_b + \text{LANDUSE}_b - 11.7303 \quad (11)$$

(where SLOPE is slope value; CURVATURE is curvature value; DRAINAGE is distance from drainage value, FAULT_b is distance from fault value, NDVI is NDVI value and ASPECT_c, LITHOLOGY_c, LANDUSE_c are logistic regression coefficient value and z_n is a parameter).

Using formula (1) and from (3) to (11), the possibility of landslide occurrence were calculated and using the possibility, eight landslide susceptibility maps were made.

4. FREQUENCY RATIO MODEL AND ITS APPLICATION

Frequency ratio approaches are based on the observed relationships between distribution of landslides and each landslide-related factor, and are used to reveal the correlation between landslide locations and the factors in the study area. Using the frequency ratio model, the spatial relationships between landslide-occurrence location and each factors contributing landslide occurrence were derived. The frequency is calculated from analysis of the relation between landslides and the considered factors. Therefore, the frequency ratios of each factor's type or range were calculated from their relationship with landslide events. In the relation analysis, the ratio is that of the area where landslides occurred to the total area, so that a value of 1 is an average value. If the value is greater than 1, it means a higher correlation, and value lower than 1 means lower correlation.

To calculate the Landslide Susceptibility Index (LSI), each factor's frequency ratio values were summed to the training area as in formula (12). The landslide susceptibility value represents the relative susceptibility to landslide occurrence. So the greater the value, the higher the susceptibility to landslide occurrence and the lower the value, the lower the susceptibility to landslide occurrence.

$$LSI = Fr_1 + Fr_2 + \ldots\ldots + Fr_n \quad (12)$$

(LSI: Landslide Susceptibility Index; Fr: Rating of each factors' type or range)

The landslide susceptibility map was made using the LSV values and for interpretation.

5. VALIDATION COMPARISON OF THE MODELS

For validation of landslide susceptibility calculation models, two basic assumptions are needed. One is that landslides are related to spatial information such as topography, soil, forest and land use, and the other is that future landslides will be precipitated by a specific impact factor such as rainfall or earthquake. In this study, the two assumptions are satisfied because the landslides were related to the spatial information and the landslides were precipitated by one cause--heavy rainfall in the study area.

The landslide susceptibility analysis result was validated using known landslide locations. Validation was performed by comparing the known landslide location data with the landslide susceptibility map. Each factor used and frequency ratio was compared. The rate curves were created and its areas of the under curve were calculated for all cases. The rate explains how well the model and factor predict the landslide. So, the area under curve in can assess the prediction accuracy qualitatively. To obtain the relative ranks for each prediction pattern, the calculated index values of all cells in the study area were sorted in descending order. Then the ordered cell values were divided into 100 classes, with accumulated 1% intervals. For example, in the case of all factor used, 90 to 100% (10%) class of the study area where the landslide susceptibility index had a higher rank could explain 45% of all the landslides. In addition, the 80 to 100% (20%) class of the study area where the landslide susceptibility index had a higher rank could explain 57% of the landslides using the logistic regression model. To compare the result quantitative, the areas under the curve were re-calculated as the total area is 1 which means perfect prediction accuracy.

So, the area under a curve can be used to assess the pre-

diction accuracy qualitatively. In the case of all factor and logistic regression model used, the area ratio was 0.786 and we could say the prediction accuracy is 78.6%. In the case of all factor and frequency ratio model used, the area ratio was 0.711 and we could say the prediction accuracy is 71.1%. In the case of slope factor and logistic regression model used, the prediction accuracy is 65.7%. Overall the case of all factor and logistic regression model used showed a higher accuracy than cases of each factor and logistic regression used and all factor and frequency ratio model used.

6. CONCLUSIONS AND DISCUSSION

The result of validation of logistic regression and frequency ratio model, the logistic regression model showed the better prediction accuracy more than 7.5%. In the case of each factor used with logistic regression, the slope used case shown the best prediction accuracy (65.7%). But, there is some difference (12.9%) from the case all used (78.7%). So, I could conclude that the case of all factors and logistic regression model used had best prediction accuracy in landslide susceptibility mapping.
The area ratio value from the effect analysis can be used to weight the relative importance of these factors, and can improve the prediction accuracy of the landslide susceptibility map. The frequency ratio model is simple, the process of input, calculation and output can be readily understood. The large amount of data can be processed in the GIS environment quickly and easily. The logistic regression model requires conversion of the data to ASCII or other formats for use in the statistical package, and later re-conversion to incorporate it into the GIS database. Moreover, it is hard to process the large amount of data in the statistical package. However, correlation of landslide and other factors can be analyzed qualitatively. The logistic regression model showed better accuracy than frequency ratio model in this study and the use of all factors show the better results. In the case of a similar statistical model (determinant analysis), the factors must have a normal distribution, and in the case of multi-regression analysis, the factors must be numerical. However, for logistical regression, the dependent variable must be input as 0 or 1, therefore the model applies well to landslide occurrence analysis.

REFERENCE

P.M.Atkinson and R.Massari, 1998, Generalized linear modeling of susceptibility to landsliding in the central Apennines, Italy. *Computer & Geosciences*, 24(4), 373-385.

F.Guzzetti, A.Carrarra, M.Cardinali and P.Reichenbach, 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy, *Geomorphology*, 31:181-216.